AUTONOMOUS NON-PROFIT ORGANIZATION
FOR HIGHER EDUCATION
"SKOLKOVO INSTITUTE OF SCIENCE AND TECHNOLOGY"

as a manuscript

Taras Khakhulin

NEW REPRESENTATIONS FOR IMAGES AND 3D SCENES

Summary of the dissertation

for the purpose of obtaining academic degree

Doctor of Philosophy in Computer Science

Academic Supervisor:

Candidate of Sciences

Victor S. Lempitsky

Moscow — 2024

Contents

Abstract **3**

List of Publications **4**

Introduction **6**

1.1 Motivation 6

1.1.1 Inductive biases in image representations for style-transfer and pixel-level generative networks 9

1.1.2 Generalizable efficient scene representation for novel view synthesis 10

1.1.3 Human priors for view and pose synthesis 11

1.2 Overview 12

1.2.1 High-resolution Image Translation for Unpaired Data 12

1.2.2 Image generators without spatial convolutions 13

1.2.3 Efficient scene representations with adaptive geometry for stereo image 14

1.2.4 Self-improving adaptive scene representation images for novel view synthesis 15

1.2.5 One-shot Mesh-based Head Avatars 16

1.2.6 One-shot High-resolution Neural Head Avatars 17

Conclusion **18**

References **20**

Abstract

The advancement of neural networks in learning from data-driven priors has opened up new possibilities in data interpretation and representation learning. While humans effortlessly convert observations into structured forms, a key facet of intelligence, artificial neural networks still rely on certain simplifications to manage this complex task. This work considers the lack of a unified approach to represent information across the diverse realms of 3D and 2D scenes.

This thesis presents different architectural inductive biases to obtain novel capabilities for images and 3D scenes. The models demonstrate new representations of spatial objects in different domains: general static 3D scenes, human avatars, and images. We investigate the new problem of translating high-resolution images from unpaired data. Then we show how to bypass the importance of architectural prior in the problem of image synthesis, particularly emphasizing positional encoding and its impacts. Employing generative adversarial networks, we demonstrate the efficient novel-view synthesis for arbitrary scenes. The pioneering scene representation allows estimate accurate reconstruction from the sparse input, and we introduce methodologies for novel view synthesis through self-improving adaptive scene representations and error correction techniques. Then close the gap between 3D and 2D methodologies, facilitating control over three-dimensional human head representations without the reliance on explicit multi-view datasets. Initially concentrating on the creation of 3D head avatars, the research investigates mesh-based proxy structures for realistic avatar generation. The thesis further extends its exploration to high-resolution synthesis of human avatars with latent control, pushing the boundaries of what is possible in this domain

List of Publications

1. Ivan Anokhin*, Pqvel Solovev*, Denis Korzhenkov*, Alexy Kharlamov*, Taras Khakhulin, Alexy Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin, ``**High-resolution daytime translation without domain labels"** in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, CORE A**.

2. Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov, ``**Image generators with conditionally-independent pixel synthesis"**, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021, CORE A**.

3. Taras Khakhulin, Denis Korzhenkov, Pavel Solovev, Gleb Sterkin, Timoteu Ardelean, and Victor Lempitsky, ``**Stereo magnification with multi-layer images"** *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.*

4. Nikita Drobyshev, Evgeny Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, Egor Zakharov. ``**MegaPortraits: One-shot Megapixel Neural Head Avatars"**. *30th ACM International Conference on Multimedia (ACMMM), 2022, CORE A*.*

5. Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, Egor Zakharov. ``**Realistic One-shot Mesh-based Head Avatars"**. *17th European Conference on Computer Vision (ECCV), 2022, CORE A*.*

6. Pavel Solovev* , Taras Khakhulin*, and Denis Korzhenkov* , ``**Self-improving multiplane-to-layer images for novel view synthesis"**. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, , CORE A.*

* denotes joint first authorship. The author of the thesis made the following contributions to the papers where he is not the first author:

- **High-resolution daytime translation without domain labels**: co-developing the training method and test pipeline, discovering metrics to evaluate the proposed approach, conducting experiments with style transfer and timelapse generation.
- **Image generators with conditionally-independent pixel synthesis**: spectral analysys of the generator, developed the test pipeline and conducting experiments to discover model properties (e.g. foveated rendering, super-resolution, inpainting), writing large portion of the text.

- **MegaPortraits: One-shot Megapixel Neural Head Avatars**: co-developing the second part of the model to produce images in high resolution, producing baselines comparison including human evaluation, conducting part of the experiments, writing a large portion of the text of the paper, preparing the figures, model schemes and method descriptions.

Introduction

The recent progress of neural networks in learning prior information from data has led us to believe that we can learn any representation from data, just as humans do. The ability to convert observations of the world into structures implicitly is a key aspect of intelligence. However, while this may be simple for humans, in artificial intelligence, we still need to introduce some shortcuts to simplify the learning process. There is currently no unified approach to achieving this, and even representing similar information in 3D and 2D scenes requires different techniques.

This thesis is presented as a line of works that proposes different architectural inductive biases to unveil novel capabilities for image processing and 3D scene representation. The manuscript introduces various approaches to learning scene representations with the priors over the content of large-scale data learned with the usage of generative adversarial networks. Furthermore, the research extends this line to encompass human-specific video data, bridging the gap between 3D and 2D approaches, and enabling control over human head representations in three dimensions without explicit multi-view datasets.

1.1 Motivation

A scene in the context of images, videos, and 3D environments is a complex construct that integrates physical attributes, environmental structure, temporal dynamics, emotional contexts, and cultural and social contexts. In recent years, significant advancements in analyzing scenes were made, including recognizing patterns [27; 39], understanding geometry [5], texture, and lighting. One of the key elements is a representation of such a scene for further processing. In the domain of images, scene representation involves interpreting and structuring visual data at the pixel level, recognizing and extracting information about objects, backgrounds, lighting, and spatial relationships. Deep learning has revolutionized this field, enabling algorithms to extract complex patterns from pixels to feature vectors, tensors, or other formats.

Images, as 2D scenes, can be accurately described as projections of 3D scenes, where the depth and spatial relationships of objects are rendered (mapped) to the flat plane. Scene representation in 3D deals with structuring data in three-dimensional space, where we have more effects that need to be represented, the simple way is to define point clouds which are sets of data points in space representing object information. Deep learning models process these point clouds to reconstruct and understand 3D scenes, such as creating detailed 3D models of urban landscapes. Another approach in 3D is using voxels, the 3D equivalents of pixels, to represent scenes in a grid-like structure for detailed and volumetric representations. Alternatively, we can define the scene as a continuous representation, where each point of the space is mapped through a function to represent the properties.

The task of constructing scene representations, in both 2D and 3D, is inherently complex due to the high dimensionality and variability of the data. Challenges arise from the need to manage diverse content, cope with varying conditions, filter out noise, and even have a realistic perception. All these factors demand substantial computational power and sophisticated algorithms to effectively understand and interpret the content. This complexity is what drives researchers in the field to continually develop and refine methods, making scene representation a dynamic and challenging domain in the intersection of computer vision and deep learning.

Any scene, whether it is 2D or 3D, can be learned from data directly, either as pixel representations or voxel representations, respectively. When we want to model certain distributions and learn priors from datasets, we are trying to efficiently extract or compress the scene information from the data. Direct learning of the image will require a separate pixel grid with N^2 parameters for each object in the data. Alternatively, the popular way in deep learning is to encode images into vectors for further analysis or even decode back to compress as much information as possible. The encoder-decoder architecture in autoencoders shows an example of inductive bias in deep learning by incorporating specific assumptions about data processing and representation, such as hierarchy of features and reconstructability. This architecture assumes that data can be understood and represented in a hierarchical manner, where the essential information can be distilled and represented in a lower-dimensional space. The meaningful patterns are those that contribute to a faithful reconstruction of the input when passed through the decoder.

Inductive bias is considered to play an important role in representing 2D and 3D scenes in machine learning [2; 17], as it can help improve efficiency and extend generalization. These biases, which are essentially the set of assumptions a learning algorithm makes

about the data, guide the learning process and enable the algorithm to efficiently interpret complex scenes, requiring fewer examples to understand underlying structures. This aspect is particularly vital in 3D scene representation where the potential data and variations are enormous. Inductive biases help the algorithm to generalize from the training data to unseen scenarios, thus predicting and understanding new scenes with greater accuracy. Moreover, they are crucial in handling ambiguities and noise in real-world data, providing a framework for interpreting such data, especially in 3D where the complexities are perspective, shadows, and occlusions. Specific biases, such as those understanding spatial relationships and depth perception, are particularly beneficial for scene representation. These biases not only reduce the computational resources required due to their structured approach to learning but also facilitate transfer learning, where a model trained on one type of scene can be adapted to another. This foundational understanding is adaptable for different types of scenes and supports complex tasks like autonomous driving, robotics, or virtual reality, where understanding and interacting with 2D/3D environments are critical.

Despite the vital importance of inductive biases for understanding the scene, in the generative tasks it can be the important piece to connect 2D and 3D, for example. We consider a popular task in generating 2D scenes - the generation of an image from the corresponding semantic map, where each pixel represents a class. The obtained results in the first attempts on the usage of generative adversarial networks already showed high quality to this task [35]. However, the results cannot be transferred to the case of a consistent generation of different views across a scene represented by a set of semantic maps (e.g. different camera views), mainly due to the high ambiguity of the input. To ensure that the table given on the stage, for example, can be generated consistently, we can add a fairly simple multi-view constraint in the underlying representation used as generator [19], which will be the inductive bias in the model. During the training of such a model, we explicitly will treat the scene as a 3D object and infuse the scene consistency by rendering it into different cameras.

Similarly, to produce realistic images and motions of humans we learn representations of digital human avatars with the existing advancements in 3D graphics to describe the head or body functionally and determine a neural rendering procedure for such scenes. Using typical assumptions or priors from 3D graphics, many tasks that seem unsolvable with conventional methods yield reasonable results in tasks involving scenes with humans. We know not only the structure of a human (e.g. skeleton) but also their basic movements, which further simplify the distribution that models need to fit [] and enable the dynamic scene understanding with the assumption of human presence on them.

We explore various ways to introduce inductive bias into the image and 3D static scene representations. As in the previous example, when training representations of 3D scenes from a set of images, we leverage the fact that we know how the scene appears from different perspectives and introduce certain constraints on the representation function and the rendering outputs. The set of constraints and assumptions is the main instrument employed in this work to facilitate the learning of novel 2D and 3D representations that are subsequently applied to various problems. We will primarily focus on the three research topics:

- unsupervised style transfer for images and 2D scene generator without direct interconnections between pixels,
- generalizable and efficient novel view synthesis
- human priors for image and video synthesis

1.1.1 Inductive biases in image representations for style-transfer and pixel-level generative networks

The problem of style transfer is one of the well-defined for 2D image representation with a number of different works in recent years we already know how effectively extract styles from the images [12] or treat content information separately [53]. Many image translation models exploit generative adversarial networks with conditional generators to inject information about the target attribute or domain [7]. Such generators are also represented with CNN-based architecture. The deep convolutional networks were shown to be highly effective in generative modeling. Stylization [13], and super-resolution [22] tasks were all shown to benefit from using the CNNs generator. Such models can be used as an efficient vector encoder from the given image [21]. Additionally, the decomposition of the image into content and style representations [29] is an efficient scheme to edit the style and obtain the desired domain transfer. Most works target the two-domain setting [58] or a setting with the fixed discrete domains [7]. We explore the image translation aims to transfer images from one domain to another, when the difference between the domains is not presented in the dataset (for instance, it can be difficult to annotate).

To overcome this limitation in this work we design training a generic image-to-image translation model on an extensive dataset of *unaligned* images without domain labels and provide an example that inductive biases from the dataset, network architecture training procedure may enable desired style transformation.

Even though the generation of realistic images started with the groundbreaking work [16], in which generative adversarial networks (GANs) were introduced, always rely on the architectural inductive bias of CNNs. CNNs inherently assume that nearby pixels are more related to each other than distant ones, a bias known as local spatial coherence and texture extraction [28; 8; 14]. Those assumptions are crucial for effectively learning features from images, as they mirrors the way objects and textures in real-world 2D scenes are typically structured, with related features often being close together. In more recent generators it was demonstrated the ability to learn the modulated kernel of convolutional decoder [20; 23; 24] to generate photorealistic plausible images. We introduce the image generator that is based on the simple MLP architecture, and each pixel is processed independently from others concerning the same noise. These models introduce new possibilities in image processing and have interesting spectral features. By predicting the color of each pixel independently and using unique coordinate encodings, CIPS showcases an innovative method of image synthesis. This model's design not only challenges conventional methodologies but also opens the horizon for more flexible and memory-efficient architectures with the same GAN paradigm.

1.1.2 Generalizable efficient scene representation for novel view synthesis

Such methods demonstrate the importance of the architectural and methodological design to succeed in some 2D tasks. When we are starting to generate 3D scenes their representation for deep learning processing becomes extremely important as mentioned above. While the CNNs are effective approaches for learning prior over 2D scenes the direct application for tasks like novel-view synthesis (NVS) [43] or even discriminative object detection [15] struggle without using architectures that take into account 3D structures (e.g. PointNet [36**?**]). The intuitive solution is to use a representation that is more friendly for CNN's inductive biases with 2D spatial locality in pixel space. Over time, various methods have been developed for creating new views. These methods can be categorized into volumetric [33; 34], mesh-based [59; 45**?** ; 18], and point-based approaches [1; 26], all of which generally require significant computational effort for rendering new views, despite their explicit usage of the 3D prior. One alternative with 2D spatial structure is depth map-based representation often sourced from stereo matching or monocular depth estimation [40]. Another key approach involves multi-layer semitransparent representations [44] has evolved with deep learning advancements to directly convert plane sweep volumes into similar representations [57], useful for interactive applications with a set of image-planes. Our work mitigates the memory requirements of the existing MPI representation [57], inspired by the 3D layered technique [40], with both geometry and color + transparency estimations done using deep

convolutional networks. For that, we introduce this new paradigm of end-to-end learning efficient representation for the scene with deep convolutional networks for stereo-imaging. We evaluate different schemes of parametrization for the network output and investigate possibilities of using GANs for more plausible images. Moreover, to alleviate per-scene learning or dataset biases we collect our dataset with thousands of static scenes.

However, the efficient on-device novel-view synthesis can be achieved with the end-to-end deep convolutional network we are still bounded by the number of input views. Next, to exted our work we are going to step further by trying to design a model that will generate a multi-layer representation for efficient scene rendering from an arbitrary number of views. The naive approach of extending the input of the network with concatenation of all possible pairs can be infeasible for a huge set of input images. The main idea of this is an attention-based neural network that fuses information from all views into a single scene representation based on the same conception from the above-mentioned multi-layer representation. To propagate as much information as we can from the input images we introduce a new technique for forward error propagation based on the concept from per-scene multi-plane images [11]. This approach unveils the ability to make a hierarchical improvement over learning-based methods for scene representation. We also show that despite the simplicity of our representation, we outperform methods that take into account view-dependencies and in theory can be more accurate on some details but in practice, it cannot achieve our quality on the forward-facing scenes.

1.1.3 Human priors for view and pose synthesis

While introduced above methods for general novel-view synthesis are the incredibly accurate framework for efficient rendering, there is a main limitation with respect to the captured data is non-rigid scenes. If something moves during capturing on the scene when we are shooting then the method based on the multi-layer concept will produce a blur in areas like this, this is caused by the nature of static data. Capturing dynamic scenes with multi-view cameras is extremely expensive and it is an incredible challenge to develop methods that support time changes for novel-view synthesis [31; 52]. To move one step beyond static scenes without using costly way of data capturing we are trying to focus on the human-specific data, more precisely on the human heads. With such prior information about the 3D nature of the data, we can learn algorithms on unsupervised monocular data [55; 41]. Priors methods mostly operate with 2D space which is a strong simplification and even small tricks about keeping information about

the 3d structure of the head [47] help to get better quality. We investigate the possibility of incorporating the geometric constraints on a deeper level by applying neural networks to render and estimate the underlying mesh-based representations. We align the paradigm of deferred neural rendering [46] with a face parametric model [30] and extend it to handle full heads without ground-truth 3D annotations (in an unsupervised way). This allows us to overcome all competitors in speed and 3D consistency for the re-enactment of human heads.

In the final part of dissertation we investigate the possibility of learning purely unsupervised models even without explicit 3D face prior. The recent success of latent avatars [4; 50] demonstrates the possibility to re-enact human portraits even without 3D knowledge at all, unfortunately, it leads to small angles for view-generation. In our approach, we introduce the explicit 3D volume that learns neural features that can be rendered into human head image warped to the desired pose. The pose and expression are disambiguated by the model with self-supervised approaches []. Furthermore, to make qualitative progress in this field we introduce an unsupervised high-resolution enhancer, inspired by HiDT from the first part of the work, that generalizes better than a naive upsampling technique since it operates on the feature level. One of the novelty of such a method is the use of a high-resolution dataset [25] without paired videos to improve quality without additional expensive data collection. This method allows us to prevail over competitors on all resolutions in self- and cross-reenactment tasks.

1.2 Overview

1.2.1 High-resolution Image Translation for Unpaired Data

In recent years image translation networks demonstrated huge abilities for modeling and editing content. Unfortunately, a carefully curated dataset with always per-pixel annotations or some subjective meta-information is necessary. The motivation for our work arises from the limitations of existing image-to-image translation methods, which typically require at least some domain labels for both training and inference. These methods have shown success in converting between two predefined paired domains (e.g. Huang et al. [2018], Isola et al. [2017], Liu et al. [2017], Zhu et al. [2017]) as well as between multiple domains (Choi et al. [2018], Lee et al. [2018, 2019], Liu et al. [2019]). However, the requirement for domain labels is a significant constraint, particularly when labels are challenging to define or collect, as is the case with varying times of day and lighting conditions. To address these challenges FUNIT [32] partially relaxes

the need for domain labels, using images from the target domain as guidance for translation in a few-shot setting. However, domain annotations remain necessary during training. Our work advances beyond this by training a multi-domain image-to-image translation model on unaligned images without domain labels, using only weak external supervision from coarse segmentation maps to boost the performance, which is not necessary based on numerical evaluation.

Our work includes two primary contributions. First, it demonstrates the feasibility of training on unpaired datasets by leveraging the internal and inductive biases of the network architecture and the dataset. Second, to ensure fine detail preservation, HiDT combines skip connections [38] with adaptive instance normalizations (AdaIN) [20]. This architectural choice is a departure from prevalent AdaIN architectures lacking skip connections.

The experimental evaluation of the HiDT model includes comparisons against several state-of-the-art baselines using objective measures and a user study. The results show the model's effectiveness in tasks such as photorealistic daytime alteration for landscape images and its potential for other multi-domain image stylization or recoloring tasks. A notable aspect of HiDT is addressing the challenge of applying image-to-image translation at high resolution, which is often computationally prohibitive. The model circumvents this by proposing an enhancement scheme that adapts the network trained at medium resolution for high-resolution images.

Overall, the HiDT model presents a method for high-resolution image translation in scenarios lacking domain labels, a significant step forward in the field of image-to-image translation. We have substantially outperformed all existing models and provided a comprehensive evaluation of the model, highlighting its strengths in dealing with high-resolution images and its versatility in various image manipulation tasks.

1.2.2 Image generators without spatial convolutions

The main building block of all generators has been a deep convolutional network, in this part we present a study of whether this is necessary and whether we can achieve the quality of modern generators without the use of directly interacting pixel features. For many years the architecutres are derived from the DCGAN [37] intuitive model for the image-decoder network with rare presence of the attention-based models [56]. However, in this work, we focus mainly on the model when the inter-pixel connection is not possibly inspired by approaches to reproduce individual scenes [34; 42].

We propose the architecture design that achieves a similar quality of generation to the state-of-the-art convolutional generator StyleGANv2 [24]. The experiments conducted to evaluate CIPS involved comparisons with state-of-the-art convolutional generators like StyleGANv2. The main building block is the periodic activation function to ingest the information about spatial positions of the pixel (e.g. place in the 2D grid).

These experiments demonstrated the CIPS architecture's similar generation quality and its unique properties, such as flexibility and efficiency in memory usage. The applications of CIPS, as indicated by these experiments, extend beyond traditional image generation tasks, offering new possibilities in fields requiring high-resolution image synthesis and manipulation without the constraints of spatial convolutions. The novel approach for image generation unlocks new applications of such networks unfeasible before. We investigate new spectral properties of our generator, memory-constrained generation, and tasks where some areas can be oversampled (e.g. super-resolution, foveated rendering).

Moreover, we show that our generators can be applied for high-resolution images with patch-based training approach when we pre-load into memory only parts of the images. Coordinate grids enable working with complex structures like cylindrical panoramas by replacing the underlying coordinate system.

1.2.3 Efficient scene representations with adaptive geometry for stereo image

Our primary motivation was the challenge of accurately capturing and representing complex scenes through stereo images, especially from unconstrained captures. Existing methods, while effective to a certain extent, fell short in dealing with complex geometries. This limitation was particularly evident in applications [40] requiring high levels of detail and realism, such as virtual reality and advanced environment reconstruction. This work aims to address these limitations by offering a more refined, multi-layered approach to scene representation, ensuring better accuracy, depth perception, and visual quality in stereo images. We are motivated by the need to fill a specific gap in the field of stereo scene representation – the ability to accurately and efficiently process complex scenes with varying depths and intricate details. By providing a multi-layered approach, StereoLayers offers a solution that is more adaptable and capable of handling the challenges inherent in stereo image processing. This method demonstrates the strong ability for memory-efficient scene representation for novel-view synthesis.

The evaluation utilized the RealEstate10k and LLFF datasets. Additionally, a new dataset named SWORD ('Scenes With Occluded Regions' Dataset) was introduced. This dataset

was specially curated to contain more diverse data with a higher prevalence of occlusions, providing a more challenging benchmark for novel view synthesis methods. The SWORD dataset was found to be instrumental in training more powerful models despite its smaller size compared to RealEstate10k

Overall, the outcomes from the experiments demonstrate the effectiveness of the StereoLayers approach in providing high-quality, adaptive, and efficient scene representations suitable for novel view synthesis, particularly in challenging scenarios involving complex geometries and occlusions. We compare our approach against existing methods like StereoMag [57] (using regularly spaced layers) and IBRNet [48] - a more recent system modeling the radiance field of the scene, and it was observed that the scene-adaptive geometry utilized in the StereoLayers approach resulted in better quality for novel view synthesis compared to non-adaptive geometry methods.

1.2.4 Self-improving adaptive scene representation images for novel view synthesis

While the efficient novel-view synthesis system can be designed for stereo input, it is a clear limitation to surpass the existing method. In this work, we focus on improving the quality of the multi-layer representation and unlocking the possibility to estimate such representation from the arbitrary number of images. Additionally, we develop a system that not only produces high-quality scene representations but also continuously improves its performance through learning and adaptation with feed-forward error propagation inspired by DeepView [11].

To achieve accurate scene reconstruction we begin with the prediction of a low-resolution fronto-parallel semi-trasparent planes [44; 57] that contains the main part of the scene geometry. As it was shown in our previous work the necessity of dense geometry is redundant and the scene can be represented more accurately with a much lower number of parameters. Based on this idea we define the conversation of the predicted planes into deformable layers in an end-to-end manner. This transformation is a key aspect of the SIMPLI method, enabling more flexible and accurate scene representations. A significant feature of SIMPLI is its feed-forward refinement procedure, which corrects the estimated representation by aggregating information from input views. This self-improving mechanism ensures that the method continually enhances its accuracy and quality of representation.

Unlike many other methods, SIMPLI does not require any fine-tuning when processing new scenes. This makes the method more practical and versatile for different applications. Overall, we demonstrate an efficient method with all the benefits from StereoLayers and more accurate scene estimation from any number of input views, based on comparison with existing state-of-the-art methods our representation takes the best from all of them and produces on-par quality for different domains. The one limitation that we ignore in this study of multi-layer geometry is dynamic scenes, the naive approach here will be data-driven video interpolation between frames [3]. Another existing drawback is view-dependent information that we ignore to achieve real-time on-device rendering [51].

1.2.5 One-shot Mesh-based Head Avatars

While standard image-to-image translation methods were able to address the problem of avatar creation for individual subjects [47; 55] or even portraits, such approaches still require a large amount of training data and difficult to train [49]. Inspired by the efficient integration of texture and adaptive geometry in the previous chapter, and recent success in neural rendering [46; 34] we tackle the problem of synthesizing personalized avatars. One limitation of existing methods, trained on large monocular video corpuses, is a small field of view and limited or intractable control of emotion and pose. This cannot be overcome without prior knowledge of 3D head geometry.

The method behind the Realistic One-shot Mesh-based Head Avatars (ROME) integrates neural networks to render photorealistic 3D human head models from just a single photograph. Our system employs the DECA [10] for accurate face and 3D pose estimation. This crucial phase ensures that the facial features are reconstructed with high fidelity, laying a foundation for the subsequent steps. The next step involves the reconstruction of the head mesh, here we predict personalized mesh for non-facial regions, extending the reconstruction beyond the facial area to include the whole head. The next step involves rendering the personalized head mesh using the estimated neural texture [46; 47]. This step is where the head avatar gains its photorealistic quality, bridging the gap between a 2D image and a 3D model.

The ROME system represents a significant advancement in creating realistic, one-shot mesh-based human head avatars. It outperforms existing methods in head geometry recovery and rendering quality, especially in cross- and self-driving scenarios. The system's ability to work without direct 3D supervision and its compatibility with existing FLAME head models are particularly notable. Future work might explore addressing current limitations such as the handling of long hair and scalability to various scales.

Despite its advancements, ROME has limitations, including challenges in modeling certain complex features like long hair and clothing, and a bias towards frontal views due to dataset limitations. Moreover, due to the lack of details in the geometry [10], we cannot model high-resolution images using the same approaches without any tricks.

1.2.6 One-shot High-resolution Neural Head Avatars

Having direct control of the model is very convenient for several VR applications. However, such models have a very strong dependence on the underlying geometry model (e.g. FLAME [9]) that leads to restriction of emotion and pose space. Despite the existing fast latent image-to-image models for human heads [55; 54] we introduce the 3D inductive bias into the latent space. First, we define the canonical information for each person that can be extracted from arbitrary images, and then, disentangle the emotion and pose information [4] to produce neutral canonical features using the self-supervised learning of latent motion. Finally, we present a novel approach to significantly enhance the quality of a single-image reenactment module by integrating high-resolution image dataset [24] into the training process similar to the first chapter. The main idea is to preserve features for the identity and train a model in a self-supervised way [6] when we do not have direct access to the re-posed images. This new training paradigm allows us to achieve a high-resolution one-shot reenactment. The experiments show that the proposed method is overall on par in self-driving mode when the emotion is extracted from the same identity as the input image, and succeed all existing methods in the cross-driving scenario.

Conclusion

In this work, several methods for the synthesis of images and 3D scenes have been introduced. The main theme that unifies the work is the direct incorporation of inductive assumptions about the data, the task, and the world into the synthesis methods themselves. This approach enables the learning of representations directly from the data, which are then applicable in further contexts.

The key contributions of this work are as follows:

1. **Image Translation**: High-resolution image translation for unpaired data has been established as a crucial area, enhancing traditional style transfer with new techniques.
2. **Image Generation**: A new approach for image generation has been introduced, which departs from traditional convolutional biases and enables advanced capabilities like content recognition, foveated rendering, and image inpainting.
3. **View Synthesis**: Techniques for efficient view synthesis have been proposed, utilizing a sparse set of images to reconstruct 3D scenes, which reduces computational load and increases adaptability to different scenes.
4. **Head Avatars**: Progress has been made on one-shot head avatars in two main directions: generative avatars that learn the 3D head and its motion, and mesh-based avatars that apply a facial prior directly.

Potential future directions include:

- Extending novel-view synthesis to dynamic scenes, tackling the complexities introduced by motion.
- Unifying the best practices from generative and mesh-based avatar approaches to enhance realism.

- Expanding the methodologies to encompass full human bodies, despite the challenges posed by sparse data inputs.
- Bridging classical rendering with learned motion and appearance to create realistic human rendering systems.

The generative modeling approach remains crucial for achieving highly realistic synthesis, with diffusion-based models suggested for generating more diverse and realistic samples.

In conclusion, the thesis culminates with the exploration of one-shot, high-resolution neural head avatars, interlacing the themes of high-resolution imaging and digital avatar creation, and showcasing advanced capabilities in generating detailed and lifelike digital human models.

Bibliography

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020.

[2] Peter Battaglia, Jessica Blake Chandler Hamrick, Victor Bapst, Alvaro Sanchez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andy Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Jayne Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. In *NeurIPS*, 2018.

[3] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. In *ACM TOG*, 2020.

[4] Egor Burkov, I. Pasechnik, Artur Grigorev, and Victor S. Lempitsky. Neural head reenactment with latent pose descriptors. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13783--13792, 2020.

[5] Angel Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. arXiv, 2015.

[6] Shuaijun Chen, Zhen Han, Enyan Dai, Xu Jia, Ziluan Liu, Xing Liu, Xueyi Zou, Chunjing Xu, Jianzhuang Liu, and Qi Tian. Unsupervised image super-resolution with an indirect supervised path. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[7] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In *2018*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8789--8797, June 2018.

[8] Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. In *ICLR*, 2017.

[9] Andrew Feng, Dan Casas, and Ari Shapiro. Avatar reshaping and automatic rigging using a deformable model. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pages 57--64. ACM, 2015.

[10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40:1 -- 13, 2020.

[11] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Styles Overbeck, Noah Snavely, and Richard Tucker. Deepview: High-quality view synthesis by learned gradient descent. In *CVPR*, 2019.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414--2423, June 2016.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, pages 2414--2423, 2016.

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[15] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *CVPR*, 2022.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, pages 2672--2680, 2014.

[17] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478, 2020.

[18] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. In *ACM TOG*, 2018.

[19] Hsin-Ping Huang, Hung-Yu Tseng, Hsin-Ying Lee, and Jia-Bin Huang. Semantic view synthesis. In *ECCV*, 2020.

[20] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510--1519, Oct 2017.

[21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-Image Translation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision -- ECCV 2018*, pages 179--196, Cham, 2018. Springer International Publishing.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, pages 694--711, 2016.

[23] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4396--4405, 2019.

[24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proc. CVPR*, pages 8107--8116, 2020.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2019.

[26] Leif Kobbelt and Mario Botsch. A survey of point-based techniques in computer graphics. *Computers & Graphics*, 2004.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 -- 90, 2012.

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278--2324, 1998.

[29] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: diverse image-to-image translation via disentangled representations. *CoRR*, abs/1905.01270, 2019.

[30] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36:1 -- 17, 2017.

[31] Tianye Li, Miroslava Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, S. Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022.

[32] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[33] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. In *ACM TOG*, 2019.

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proc. ECCV*, pages 405--421, Cham, 2020. Springer International Publishing.

[35] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2019.

[36] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211--252, 2015.

[40] M. L. Shih, S. Y. Su, J. Kopf, and J. B. Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020.

[41] Aliaksandr Siarohin, Stéphane Lathuilière, S. Tulyakov, Elisa Ricci, and N. Sebe. First order motion model for image animation. *ArXiv*, abs/2003.00196, 2019.

[42] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Proc. NeurIPS*. Curran Associates, Inc., 2020.

[43] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019.

[44] Richard Szeliski and Polina Golland. Stereo matching with transparency and matting. 1999.

[45] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. In *ACM TOG*, 2019.

[46] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[47] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: real-time face capture and reenactment of rgb videos. *ArXiv*, abs/2007.14808, 2019.

[48] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*, 2021.

[49] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[50] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10034--10044, 2021.

[51] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. NeX: Real-Time View Synthesis With Neural Basis Expansion. In *CVPR*, 2021.

[52] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022.

[53] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic Style Transfer via Wavelet Transforms. *arXiv:1903.09760 [cs]*, March 2019. arXiv: 1903.09760.

[54] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor S. Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020.

[55] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. 2019.

[56] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019.

[57] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM TOG*, 2018.

[58] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242--2251, October 2017.

[59] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. In *ACM TOG*, 2004.